

一种基于稀疏主成分的基因表达数据特征提取方法

沈宁敏 李 静 周培云 庄 毅

(南京航空航天大学计算机科学与技术学院 南京 210016)

摘 要 聚类已成为基因表达数据的一种前沿分析方法,通过基因类别的划分可以较快速地发现病变细胞,以实现疾病的诊断。然而,高维、小样本的数据特点使得原始采集的基因表达数据具有大量的冗余与干扰信息,直接聚类会使得算法运行时间长,分析结果精度低。主成分分析是一种经典的数据降维方法,在保持方差最大的情况下,将高维数据映射到低维空间。但负载因子的非零特性使得主成分不具有强解释能力。提出基于截断幂的稀疏主成分分析方法对基因表达数据进行特征提取,并结合 K-means 方法对稀疏提取的特征基因数据进行聚类分析。最后,利用 3 个公开的基因数据集进行实验分析,验证了所提出的特征提取方法可提高基因表达数据聚类的精确性与高效性。

关键词 基因表达数据,负载因子,截断幂,稀疏主成分分析,特征提取

中图法分类号 TP399 文献标识码 A

Feature Extraction Method Based on Sparse Principal Components for Gene Expression Data

SHEN Ning-min LI Jing ZHOU Pei-yun ZHUANG Yi

(College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China)

Abstract Cluster analysis is a popular method for gene expression data, which can be used for finding cancer cell so that the diseases can be diagnosed accurately and rapidly through the gene class label. However, more attributes and less samples will produce a mass of redundant or disturbed information, resulting in the decline of the accuracy of the direct clustering in high dimensional data. Principal Component Analysis (PCA) is a classical method for dimension reduction which can transform high dimension data into low space under maintaining maximal variance. The shortcoming of PCA is the lack of strong interpretation for loadings that have no characteristic of sparsity. In this paper, a sparse PCA method based on Truncated Power was applied into the feature extraction for gene expression data, then the sparse PCA was fed into K-means process for clustering. Finally, the experimental results on Colon cancer, leukemia and lung cancer three typical gene datasets verify that the sparse gene data can improve the efficiency and accuracy on clustering.

Keywords Gene expression data, Loadings, Truncated power, Sparse principal component analysis, Feature extraction

1 引言

伴随计算与生物技术的快速发展, DNA 微阵列芯片技术可快速、准确地评估基因表达水平。基因表达是指生物中的信使 RNA 经过转录与翻译生成具有活性或病变的蛋白质活性分子。通过分析基因表达数据可寻找出具有生物标记的基因,以区分正常细胞与癌细胞,提高对疾病的确诊几率,对人类正常生活产生深远影响。近些年,对基因表达芯片数据进行系统性分析已成为数据挖掘、机器学习与模式识别等领域的热门研究课题^[1-3]。但基因表达数据具有基因多、样本少的数据特点。由于高维分布的数据存在大量冗余或干扰信息,并且在基因表达过程中多数基因是密切相关的,因此直接对基因数据进行聚类分析会导致精确率低,耗时也长。针对上述问题,目前有各种不同的数据降维方法被用于基因数据的特征选择与模型识别中^[4-6]。独立成分分析 (Independent Component Analysis, ICA)^[7]是将高维数据分解成单个的独

立成分,通过选择基因数据的特征子集进行肿瘤分类^[8]或建立基因表达数据的线性模型^[9]。主成分分析 (Principal Components Analysis, PCA)^[10,11]是基于属性变量方差最大化原则的一种降维方法。新主成分是原有基因数据特征的线性组合,可消除数据之间的冗余与相关性,以便进行深入的基因内部关联探索^[12]或可视化的分析^[13]。尽管 ICA 与 PCA 都可对基因数据进行降维以实现特征选择,但在 ICA 中,独立成分没有主次之分,即被选择的子集基因在表达过程中的作用级别无法判断;在 PCA 中,每个基因主成分是原变量的线性组合,而负载因子 (Loadings) 的非零属性使主成分不具有可解释性,即不知具体哪些基因决定着细胞的分化。

稀疏主成分分析 (Sparse Principal Component Analysis, SPCA) 是在 PCA 的基础之上对负载因子进行稀疏化,即使得大部分负载因子为 0,从而使得主成分更具解释能力。SPCA 综合考虑了主成分的方差与负载因子的稀疏化,从而弥补了传统 PCA 的缺陷。近几年,已有不同的 SPCA 方法被提出,

本文受中央高校基本科研业务费专项资金 (NZ2013306) 资助。

沈宁敏 (1991—), 男, 硕士生, 主要研究方向为数据挖掘、并行计算, E-mail: ningminshen@163.com; 李 静 (1976—), 女, 博士, 副教授, 主要研究方向为数据挖掘、图像处理、可信软件; 周培云 (1990—), 男, 硕士生, 主要研究方向为数据挖掘、图像处理; 庄 毅 (1956—), 女, 教授, 博士生导师, 主要研究方向为分布式计算。

如基于回归的方法^[14]、半正定松弛规划^[15]与广义幂迭代^[16]等。相比之下,截断幂(Truncated Power)方法结合了幂迭代法^[17]求解矩阵特征值与特征向量,并加入变量协方差矩阵的收缩操作^[18]以确定主成分的主次,它综合考虑主成分方差、负载因子的稀疏程度与算法的运行效率,是一种更为高效的稀疏求解方法。本文设计了基于截断幂的 SPCA 方法对基因表达数据集进行稀疏特征提取,并通过 K-means 方法对特征提取后的基因表达数据进行分析,且对方法性能进行了分析。

本文第 2 节介绍稀疏主成分的原理与发展,给出了基于截断幂的稀疏 PCA 求解算法;第 3 节给出了基于稀疏 PCA 的基因表达数据的稀疏特征提取与分析方法;第 4 节通过实际数据集验证了本文所提出的方法的性能,最后总结了本文所做工作及将来的研究方向。

2 稀疏主成分分析

2.1 概述

给定数据矩阵 $A \in R^{n \times p}$, n 为样本个数, p 为属性个数,对数据集进行正则化处理之后得到变量协方差矩阵 $\Sigma = A^T A$, 则稀疏主成分分析的求解公式为:

$$x^* = \arg \max x^T \Sigma x \quad (1)$$

满足 $x^T x = 1, x \in R^n$ 且 $\text{card}(x) \leq k, k(k > 0)$ 为负载因子的非零个数,控制稀疏程度的调优参数, x^* 为所求的最优负载因子。该公式模型的求解是非凸优化问题,是 NP 难解问题,其求解的方式是利用松弛近似等方法对公式模型求近似最优解。

SPCA 的求解已有多种不同类型的方法, Cadim (1995)^[19] 将一些绝对值较小的变量系数近似为 0,并在此基础上进一步利用坐标旋转技术分析主成分与变量之间的相关性,找出主成分在变量子空间下的近似稀疏线性组合。Vines (2000)^[20] 在二维子空间中对主成分近似地转换以限定变量系数为 -1、0、1 这 3 个整数。Tibshirani (2003)^[21] 从系数向量的整体出发,结合软阈值将主成分的系数向量范数设为小于某一个限定的数。Zou (2006)^[13] 受 Lasso 问题的启发,将稀疏负载因子的求解问题看作线性模型中的变量选择问题,提出一种弹性网(Elastic-net)的惩罚项函数的回归模型,它可以很好地解决样本个数远远大于变量个数的数据集问题。Journee (2010)^[15] 从数学与主成分并行提取的角度出发,对原有的稀疏公式模型进行近似优化推导,以 ℓ_0 和 ℓ_1 两种惩罚函数为代表,形成 4 种不用应用场景的最优公式求解模型。

2.2 截断幂

截断幂(Truncated Power)方法是 Yuan (2013)^[16] 以非凸优化问题的近似求解为动机,从矩阵特征值求解的问题出发,应用截断的思想提出的一种高效、稳定的 SPCA 求解方法,其算法如表 1 所列。算法的主要过程包含幂迭代与截断稀疏,幂迭代是一种求解矩阵最大特征值与对应特征向量的方法,而截断是指通过定义的算子对所求的特征向量进行稀疏化处理。截断幂 SPCA 算法中对负载因子稀疏的算子定义如下:

$$\text{Truncated}(x, F) = \begin{cases} [x]_i, & i \in F \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

其中, x 为负载因子向量, F 为截断下标集合。当一个主成分的负载因子稀疏以后,根据式(3)对协方差矩阵进行收缩操作以保证提取的主成分具有主次之分,即所有主成分的方差是呈下降趋势。

$$\Sigma' = (I_{p \times p} - x^* x^{*T}) \Sigma (I_{p \times p} - x^* x^{*T}) \quad (3)$$

表 1 截断幂算法

算法: 截断幂稀疏主成分分析	
输入:	数据集矩阵 $A \in R^{n \times p}$, 包含 n 个样本, p 个变量; m 个主成分; 协方差矩阵 $\Sigma \in R^{p \times p}$; 负载因子的基数向量 $k \in \{k_1, k_2, \dots, k_m\}$
输出:	负载因子向量 $x = \{x^1, x^2, \dots, x^m\}$
for	$i = 1 : m$
{	
	初始化负载因子向量 $x_0^i = \vec{1}$, 迭代次数 $t = 1$;
do{	
	1. 使用幂迭代法计算 $x_t^{i'} = \Sigma x_{t-1}^i / \ \Sigma x_{t-1}^i\ $;
	2. 根据 $x_t^{i'}$ 向量中的具体值按从大到小的顺序提取前 k 个下标组成集合 F_t^i ;
	3. 使用截断算子 $x_t^{i'} = \text{truncated}[x_t^{i'}, F_t^i]$ 对 $x_t^{i'}$ 进行截断稀疏;
	4. 正则化负载因子向量 $x_t^i = x_t^{i'} / \ x_t^{i'}\ $, 迭代次数 $t = t + 1$
}while{	直到收敛};
	协方差矩阵进行收缩;
	$\Sigma = (I - x^i x^{iT}) \Sigma (I - x^i x^{iT}), I \in R^{p \times p}$
}	

截断幂 SPCA 算法的输入是初始的数据集 A 、主成分的个数 m 和负载因子的非零个数即基数(Cardinality)向量 k , 输出为所求的稀疏负载因子向量组合 $x = \{x^1, x^2, \dots, x^m\}$ 。内层循环利用截断与幂结合方法求出单个主成分的负载因子,外层利用对协方差矩阵收缩求后续的主成分。利用截断与收缩反复迭代可以在最大化方差的同时保证基数的最小化。

3 本文方法

本文所提出的基因表达数据稀疏特征分析方法主要包括基因数据的预处理、基于稀疏主成分分析的基因选择及 K-means 聚类的基因分析 3 个步骤,其框架如图 1 所示。

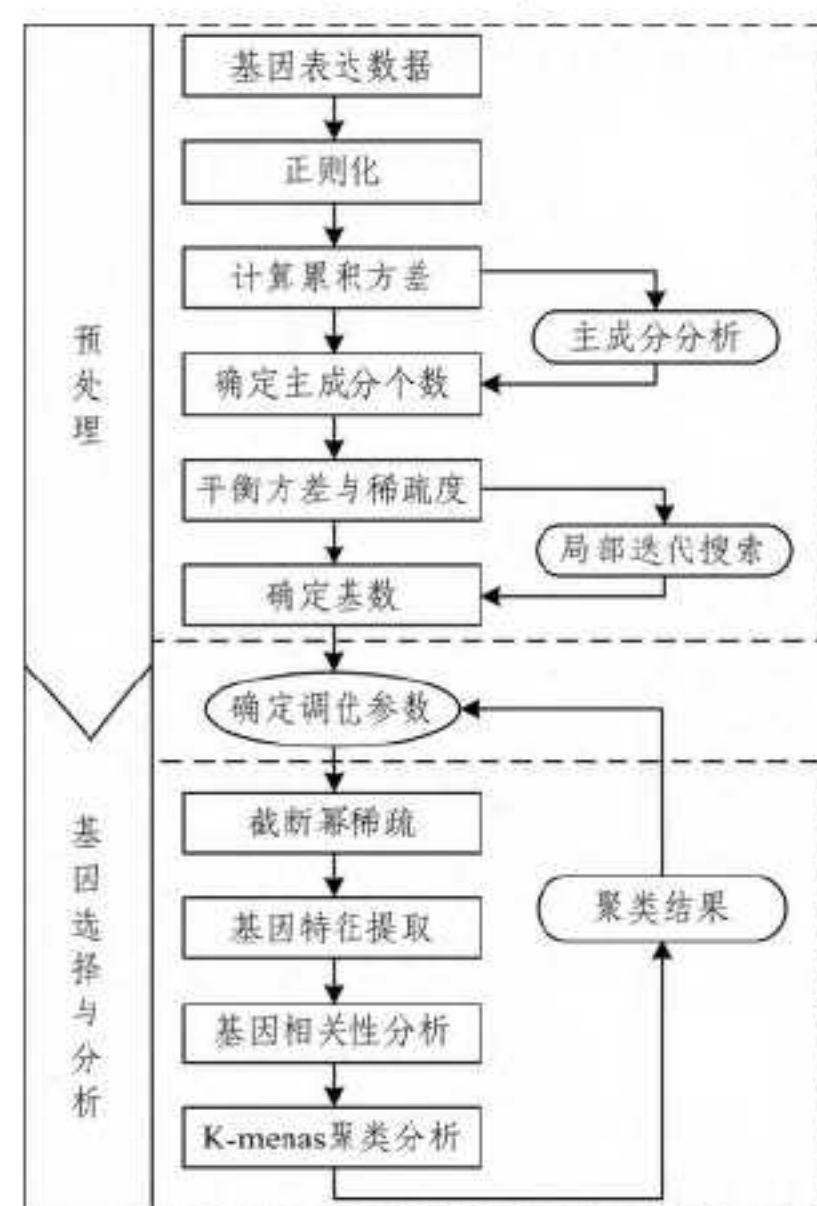


图 1 基因数据稀疏特征分析框架

DNA 微阵列在采集基因表达数据时,由于实验过程中可能会受到一些外界环境或偶然因素的干扰,造成生成的基因数据集不具有统一性,需对原始的数据集进行正则化处理,目的是消除阵列芯片在采集数据时所造成的系统误差,使得多次测量的基因表达数据分布均匀。应用截断幂方法对基因表达数据进行稀疏特征提取需提供两个调优参数,即主成分的个数和负载因子的基数,这两个参数分别利用主成分的方差分析与基数值值的迭代搜索来确定。当初始的调优参数确定

后,对预处理后的基因表达数据进行稀疏处理,对提取出的基因进行相关性分析和 K-means 聚类分析,并将聚类精度的结果反馈到调优参数的设置上,以达到最优结果。

3.1 预处理

基因表达数据集可通过一个矩阵 $A \in R^{n \times p}$ 来描述,其中矩阵元素 $a(i, j)$ 代表第 j 个基因在第 i 次测量中的表达水平。基因数据预处理的步骤如下:

Step1 对原始样本集做正则化处理,使得每个基因的样本值均值为 0,方差为 1;计算基因之间的协方差矩阵 $\Sigma(p \times p)$,其中 $\Sigma(i, j)$ 表示第 i 个基因与第 j 个基因之间的相关性,值越大,说明两者在基因表达过程中关联程度越高。

Step2 利用主成分分析法对协方差矩阵 Σ 进行特征值分解,将所求的特征值按从大到小的顺序进行排序, $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$,每个主成分的贡献率由对应特征值的大小决定,同时,基因的表达特征水平与主成分的累积方差贡献率是成正比的;计算前 m 个主成分的累积方差,当其值大于一定的阈值 ϵ 时,即可确定主成分的个数,累积方差通过式(4)计算。

$$\sum_{i=1}^m \lambda_i / \sum_{j=1}^p \lambda_j \geq \epsilon \quad (4)$$

Step3 稀疏主成分分析中最主要的问题是如何确定负载因子的基数。本文设计一种本地局部的迭代搜索方法来平衡主成分的方差与负载因子的稀疏度。首先,给定一个公共误差阈值 δ ,则第 i 个稀疏主成分的方差范围为 $(pev_i - \delta, pev_i + \delta)$, pev_i 为第 i 个主成分的方差;然后根据 $pev(i) = R_{i1}^2 / tr(A^T A)$ 与 $Z_i = Q_i R_i$,在给定 $Q_i (i=1, 2)$ 初始值的条件下,负载因子 $l_i (i=1, 2)$ 通过 $l_i = Z_i A^{-1}$,可以从各个主成分的基数上限 φ 与下限 ϕ 近似得到。最后,在给定的近似范围 $[\phi, \varphi]$ 内,对每个主成分进行本地局部搜索,使得第 t 次迭代过程中满足 $|pev(t) - pev(i)| < \xi$,则对应的基数值可以被确定。

3.2 基因选择

对数据进行预处理之后,主成分的个数与负载因子的基数这两个调优参数都已初步确定,使用表 1 列出的截断幂稀疏算法可以获取稀疏后的主成分与对应的负载因子。负载因子的大小可近似看作在基因表达过程中某个基因表达水平的权值比重,根据其大小可以列出一些重要的基因并分析它们之间的相关性,相关性越大,说明稀疏基因特征的提取就越精确。基因 x 和基因 y 之间的相关性为:

$$r(x, y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 (y_i - \bar{y})^2}} \quad (5)$$

其中, $\bar{x}$ 与 $\bar{y}$ 分别是所有样本下基因 x 与基因 y 的均值。稀疏基因主成分的特征提取通过 $Z = A * loadings$ 计算, Z 最后的维度是 n 个样本, m 个基因。

3.3 基因聚类

稀疏提取的主成分将通过聚类分析来验证其是否能代表真实的基因表达水平。K-means 聚类是一种典型的聚类方法,其主要思想为:事先给定聚类的类别 k 与基因主成分 $Z = \{z_1, z_2, \dots, z_m\}$, k 个初始的聚类中心 $U = \{u_1, u_2, \dots, u_k\}$ 将通过随机的方式进行选择,然后根据样本与聚类中心的欧氏距离进行比较,重新确定新的聚类中心,循环操作,直到聚类中心收敛为止。对于给定一个聚类簇 c_i ,一个样本是否属于该簇,由其与中心之间的欧氏距离决定, $c_i = \{z_i | \forall f, d(\vec{z}_i, \vec{u}_j) < d(\vec{z}_j, \vec{u}_i)\}$ 。欧氏距离的计算公式如下:

$$d(\vec{z}_i, \vec{u}_j) = \sqrt{(\vec{z}_i - \vec{u}_j)^2} \quad (6)$$

其中, $\vec{z}_i$ 为第 i 个样本, $\vec{u}_j$ 为第 j 个聚类中心, $d(i, j)$ 为第 i 个基因与第 j 个基因的距离。然后所有簇的均值 $\vec{u}_j$ 将被重新更新,更新公式为:

$$\vec{u}_j = \frac{1}{|c_j|} \sum_{\vec{z}_i \in c_j} \vec{z}_i \quad (7)$$

4 实验与讨论

为了评估截断幂稀疏主成分求解方法对基因特征分析的性能,本文将上述所提出的方法应用到结肠癌、白血病与肺癌 3 个典型的基因数据集。实验平台为 PC 机 (Windows, Inter (R) Core(TM) i5-3470 CPU@3.20GHz),所使用的软件为 Matlab2010b。实验中对所使用的初始化参数分别进行了设定:预处理步骤中累积方差比例参数 $\epsilon = 0.7$,稀疏主成分方差范围参数 $\delta = 0.3$,主成分最小容错参数 ξ 是一个变化的常数,其大小与主成分的次序有关。两个数据集的聚类类别个数分别为 2 和 3。

4.1 数据集

本文所使用的数据集为公开的基因表达数据集,分别为结肠癌 (Colon cancer)、白血病 (Leukemia) 及肺癌 (Lung cancer)。3 个数据集的样本类别及基因个数如表 2 所列,实验中通过样本聚类的准确率来衡量截断幂稀疏特征提取方法的有效性。

表 2 实验数据集分布

数据集	样本	基因	类别			
结肠癌	62	2000	Tumor		Normal	
			22		40	
白血病	38	5000	ALL-B	ALL-T	AML	
			19	8	11	
肺癌	197	1000	AD	NL	SQ	COID
			159	17	21	19

4.2 实验结果与分析

4.2.1 确定 PCA 主成分个数

在使用截断幂方法对基因数据进行稀疏特征提取之前,需确定 PCA 提取主成分的个数。PCA 中所提取的主成分的方差大小随着主成分提取的顺序越来越小,当前 m 个主成分的累积方差达到特定比例时,余下的数据可以看作是冗余信息。

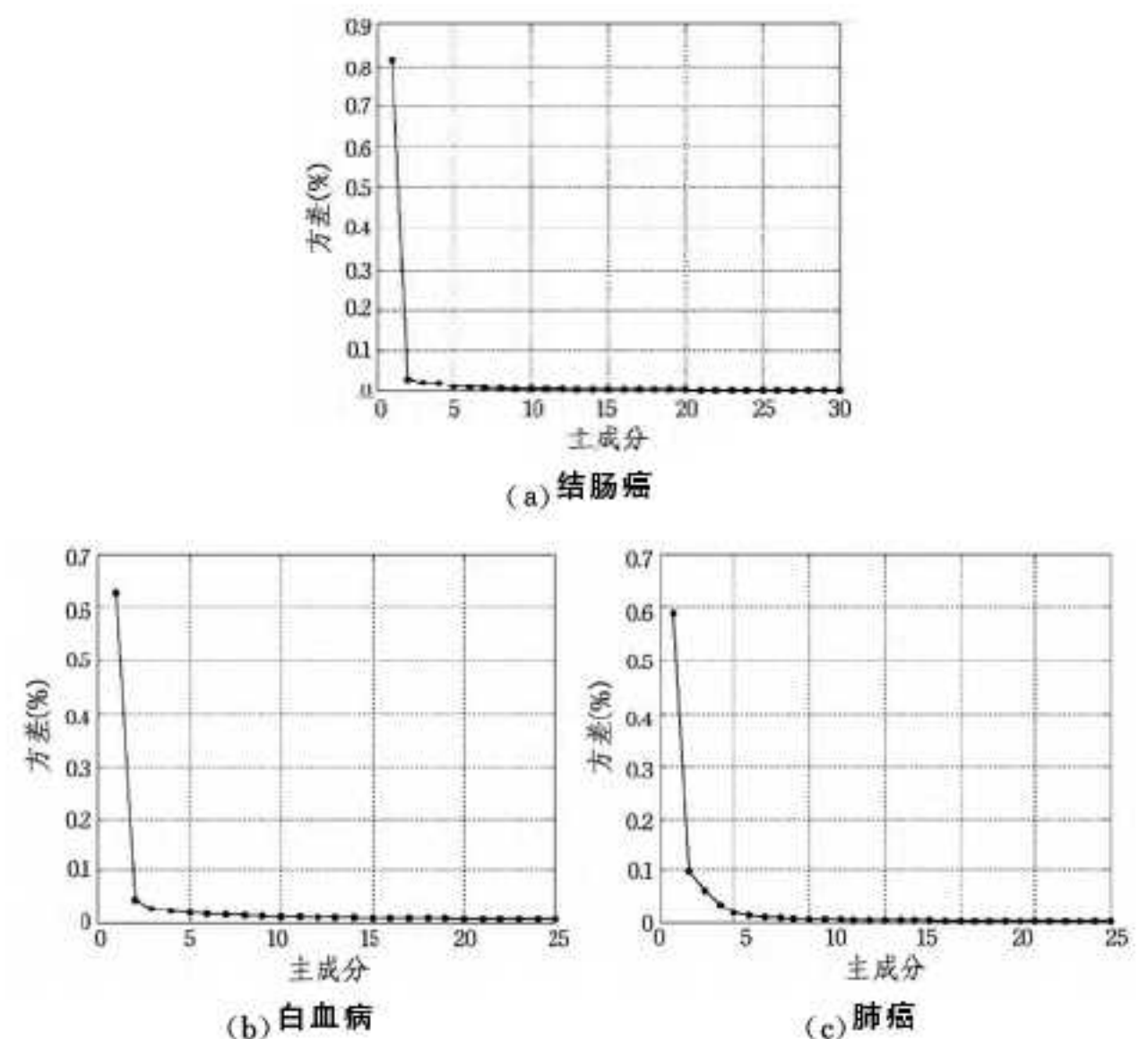


图 2 基因主成分的方差与个数的关系

图2给出了3个基因数据集分别在前30与25个主成分的方差变化趋势。由表3可以看出,前3个主成分的累积方差与主成分的整体方差比例已到达实验预定的参数 $\epsilon=0.7$,因此主成分个数确定为3。

表3 基因数据的前3个主成分的方差与累积方差

数据集	方差(%)			累积方差(%)
	PC1	PC2	PC3	
结肠癌	81.50	3.13	2.22	86.85
白血病	62.96	4.59	2.76	70.31
肺癌	58.82	9.75	6.20	74.77

4.2.2 局部迭代搜索法确定负载因子基数

在已有的SPCA方法中,负载因子的非零个数即基数通常是根据先验知识给定的,本文设计的局部迭代搜索方法通过稀疏主成分与非稀疏主成分之间的最小容错率近似得到负载因子的基数。当主成分个数确定以后,根据给定的容错范围对基数进行局部迭代搜索,实验结果如图3与表4所示。

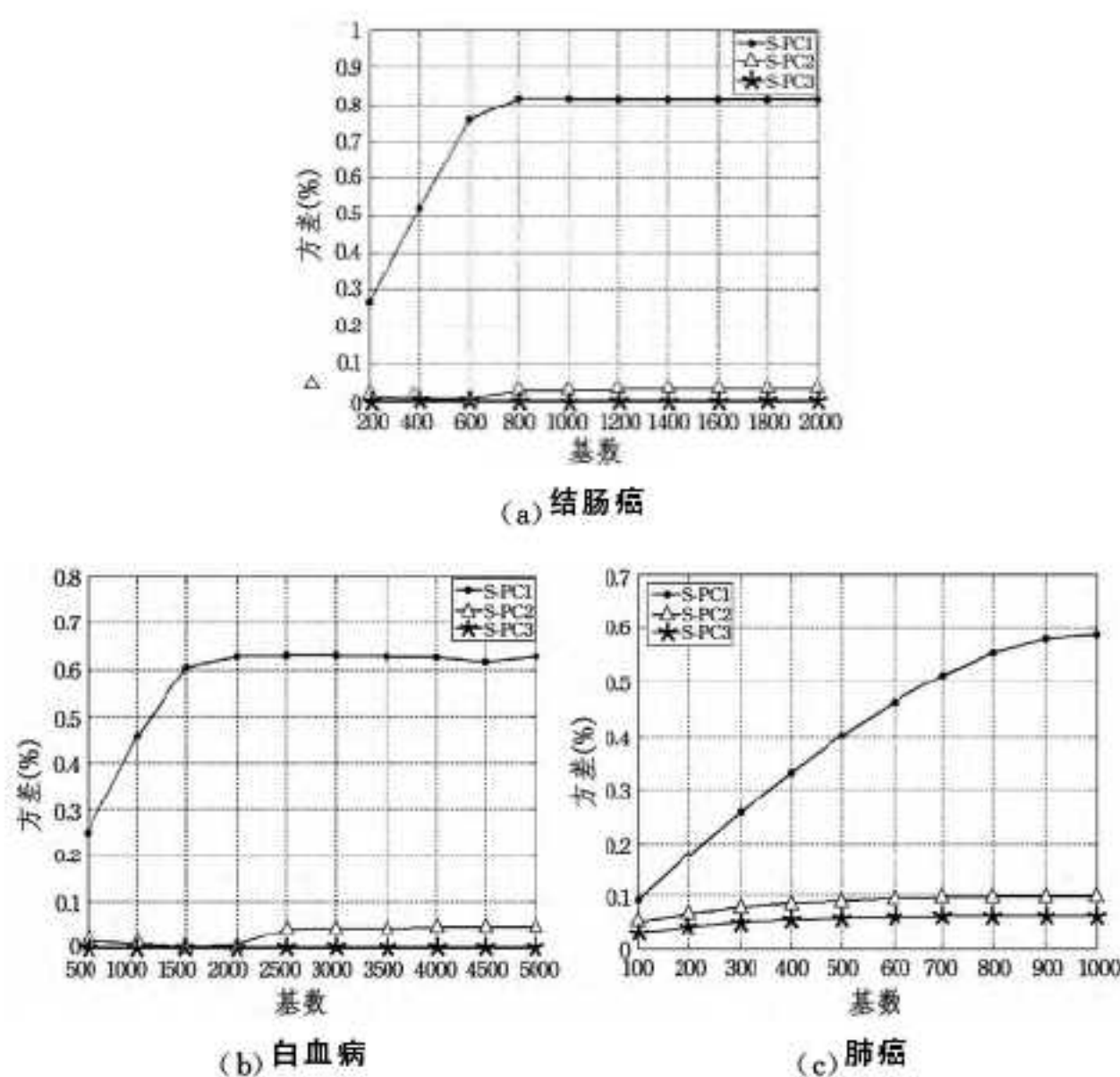


图3 前3个基因稀疏主成分方差与基数的关系

图3给出了3个基因数据集中前3个基因稀疏主成分方差与基数的关系。可以看出,随着基数增加,方差呈稳定上升趋势,当基数达到一定的范围之后,方差基本保持不变,由此可以确定主成分对应的基数,如表4所列。与表3相比,稀疏

主成分与原主成分的方差差值很小,即负载因子在稀疏化处理之后,主成分对数据的整体表示并无大变化,由此可以得出原主成分中存在一些与数据特征无关的冗余数据。

表4 基因数据的前3个稀疏主成分的方差与累积方差

数据集	方差(%)与基数(n)						累积方差 (%)
	S-PC1		S-PC2		S-PC3		
结肠癌	81.35	1200	1.30	800	0.97	400	83.80
白血病	60.85	2500	2.89	2500	0.91	1000	64.65
肺癌	59.05	900	9.20	500	5.03	300	73.28

4.2.3 重要基因相关性分析

上述基因表达数据集基因个数虽然很多,但生成活性蛋白质分子的过程仅由紧密关联的少数基因决定。在本文方法中基因表达的整体水平可以依据稀疏主成分的负载因子,当新主成分由强相关性的稀疏基因进行线性组合时,即代表主成分对整体数据更具有代表性,这与后续样本数据分类或聚类是紧密关联的。本实验分析上述3个基因数据集前10个强表达基因之间的相关性,如图4所示,通过SPCA方法提取的稀疏特征基因的相关性均大于0.6,关系密切,同时表5给出了3个基因数据集强表达基因的权重及其描述。

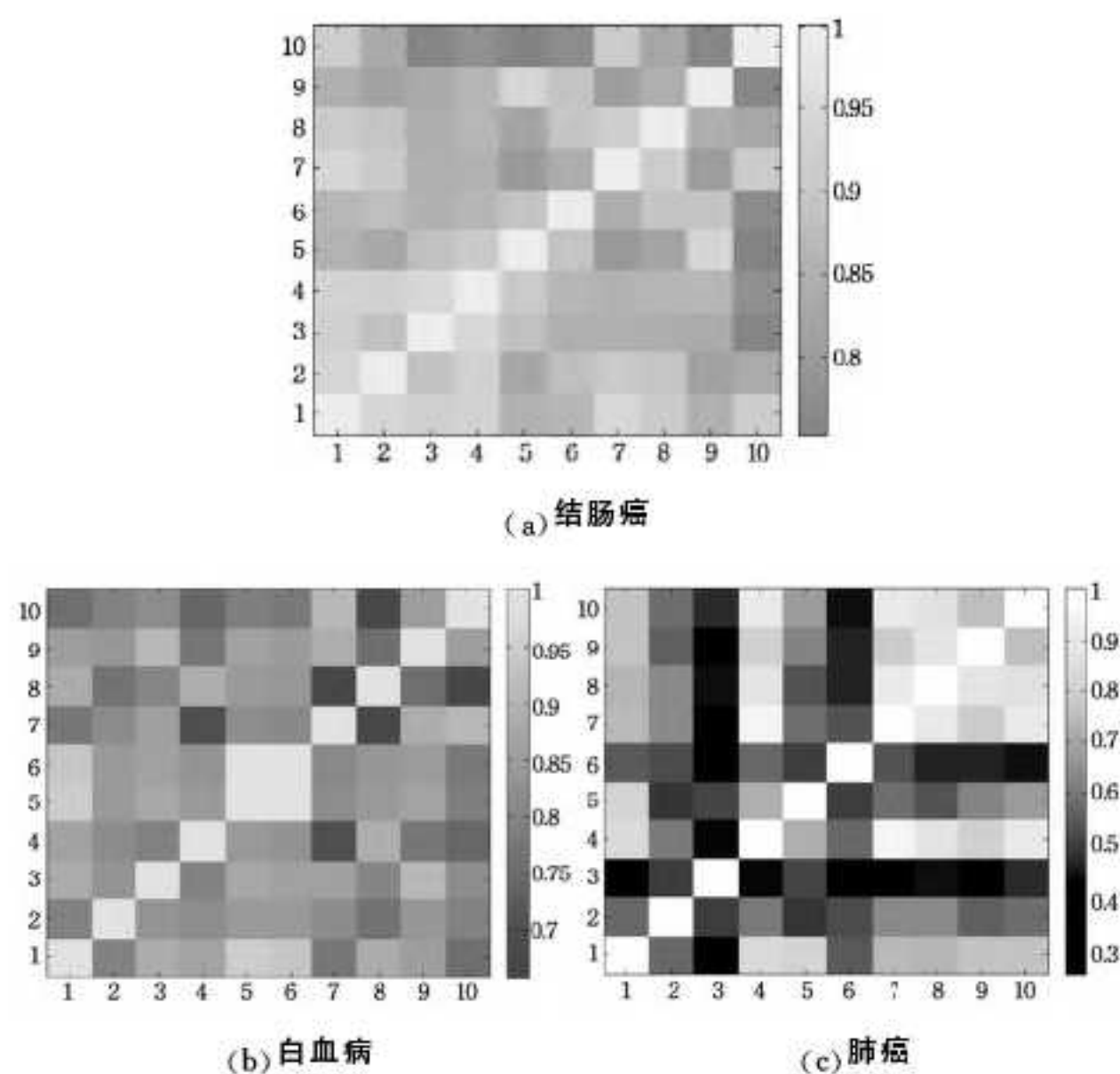


图4 前10个重要基因的相关性

表5 前10个重要基因的基本信息

(a) 结肠癌

序号	ID	表达权重	描述
1	415	0.1220	"ACTIN, AORTIC SMOOTH MUSCLE(HUMAN)"
2	1967	0.1196	"RAN GTPASE ACTIVATING PROTEIN 1(Mus musculus)"
3	1423	0.1194	M17446-s-at "FGF4 Fibroblast growth factor 4(heparin secretory transforming protein 1,Kaposi sarcoma oncogene)"
4	822	0.1089	"MYOSIN REGULATORY LIGHT CHAIN 2, SMOOTH MUSCLE ISOFORM(HUMAN); contains element TAR1 repetitive element"
5	1494	0.1078	"H. sapiens mRNA for hevin like protein"
6	897	0.1062	"COMPLEMENT FACTOR D PRECURSOR(Homo sapiens)"
7	1285	0.1054	"Human mRNA for erythrocyte membrane sialoglycoprotein beta(glycophorin C)"
8	1272	0.1052	"TRISTETRAPROLINE(HUMAN)"
9	1843	0.1046	"GELSOLIN PRECURSOR, PLASMA(HUMAN)"
10	1545	0.1042	"METALLOPROTEINASE INHIBITOR 3 PRECURSOR(Gallus gallus)"

(b) 白血病

序号	ID	表达权重	描述
1	1356	0.0688	Y10256-at Serine/threonine protein kinase, NIK
2	1243	0.0666	X03794-s-at HOXB5 Homeo box B5(2.1 protein)
3	438	0.0665	M17446-s-at FGF4 Fibroblast growth factor 4(heparin secretory transforming protein 1,Kaposi sarcoma oncogene)
4	821	0.0658	U45255-s-at Paired-box protein PAX2(PAX2) gene, exon 11 and complete cds
5	3000	0.0651	HG2160-HT2230-at Glutamate Decarboxylase 1
6	4028	0.0645	Z29066-s-at Nek2 mRNA fo protein kinase
7	4022	0.0643	D82348-at 5-aminoimidazole-4-carboxamide-1-beta-D-ribonucleoti de transformylase/inosinicase
8	3587	0.0640	X52005-at MYL4 Myosin, light polypeptide 4,alkali;atrial,embryonic
9	3640	0.0640	S62027-s-at Transducin gamma subunit [human,mRNA,408 nt]
10	254	0.0639	U10693-at MAGE-8 antigen(MAGE8) gene

(c) 肺癌

序号	ID	表达权重	描述
1	329	0.1179	transcription factor 21
2	139	0.1168	zinc finger protein 145(Kruppel-like, expressed in promyelocytic leukemia)
3	167	0.1166	aminomethyltransferase(glycine cleavage system protein T)
4	883	0.1165	TEK tyrosine kinase, endothelial(venous malformations, multiple cutaneous and mucosal)
5	105	0.1156	Cluster Incl M63438; Human Ig rearranged gamma chain mRNA, V-J-C region and complete cds / cds=(0,1049) / gb=M63438/gi=184847 / ug=Hs.156110 / len=1244
6	271	0.1144	ATP-binding cassette, sub-family A(ABC1), member 3
7	306	0.1120	tetranectin(plasminogen-binding protein)
8	824	0.1119	four and a half LIM domains 1
9	288	0.1099	A kinase(PRKA) anchor protein 2
10	323	0.1089	cadherin 5, type 2, VE-cadherin(vascular epithelium)

4.2.4 特征基因聚类

为了验证本文稀疏特征提取方法的有效性,结合 K-means 对基因稀疏主成分进行了聚类分析,同时将本文方法与传统的 ICA 及 PCA 特征选择与提取方法的精度和运行时间进行了比较。精度的计算如下:

$$AC = \frac{\sum_{i=1}^n I(i)}{n} \quad (8)$$

其中, $I(i)$ 表示第 i 个样本聚类是否正确,其值为 1 或 0, n 为基因样本的个数。本文方法与其它两种方法的聚类精度和运行时间如表 6、表 7 所列,从表中可以看出,本文提出的方法应用于基因表达数据,在可接受时间内,聚类精度优于其它两种方法。

表 6 方法聚类精度对比(%)

数据	ICA	PCA	本文方法
结肠癌	53.2	50.0	64.8
白血病	75.6	68.7	83.4
肺癌	79.3	75.2	84.6

表 7 方法运行时间对比(s)

数据	ICA	PCA	本文方法
结肠癌	5.5	15.9	2.4
白血病	6.3	228.5	36.6
肺癌	2.6	6.7	1.3

结束语 本文应用截断幂的稀疏主成分分析方法对基因表达数据进行稀疏特征提取并结合 K-means 对稀疏提取的基因主特征进行聚类分析,尝试发现一些基因在表达过程中的差异性,为肿瘤细胞的诊断等提供重要的依据。实验结果验证了本文方法的有效性。结构化稀疏基因特征数据的选择及聚类精度的提高将会在今后工作中进行。

参 考 文 献

- [1] Khobragade V P, Vinayababu A. A Classification of Microarray Gene Expression Data Using Hybrid Soft Computing Approach [J]. International Journal of Computer Science Issues (IJCSI), 2012, 9(6)
- [2] Bi X, Huang H, Matis-Mitchell S, et al. Building a classifier for identifying sentences pertaining to disease-drug relationships in
- tardive dyskinesia [C] // 2012 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2012: 1-4
- [3] Zhou X, Liu K Y, Wong S T C. Cancer classification and prediction using logistic regression with Bayesian gene selection [J]. Journal of Biomedical Informatics, 2004, 37(4): 249-259
- [4] Atallah R, Ryan J, Aeschlimann D. Incorporating Known Pathways into Gene Clustering Algorithms for Genetic Expression Data [C] // CS 229: Machine Learning Final Projects, Autumn 2013. 2013
- [5] Abraham G, Inouye M. Fast Principal Component Analysis of Large-Scale Genome-Wide Data [J]. PloS one, 2014, 9(4): e93766
- [6] Natarajan N, Dhillon I S. Inductive matrix completion for predicting gene-disease associations [J]. Bioinformatics, 2014, 30(12): i60-i68
- [7] Hyvärinen A, Karhunen J, Oja E. Independent component analysis [M]. John Wiley & Sons, 2004
- [8] Huang D S, Zheng C H. Independent component analysis-based penalized discriminant method for tumor classification using gene expression data [J]. Bioinformatics, 2006, 22(15): 1855-1862
- [9] Liebermeister W. Linear modes of gene expression determined by independent component analysis [J]. Bioinformatics, 2002, 18(1): 51-60
- [10] Smith L I. A tutorial on principal components analysis [D]. Cornell University, USA, 2002, 51: 52
- [11] Jolliffe I. Principal component analysis [M]. John Wiley & Sons, Ltd, 2005
- [12] Misra J, Schmitt W, Hwang D, et al. Interactive exploration of microarray gene expression patterns in a reduced dimensional space [J]. Genome research, 2002, 12(7): 1112-1120
- [13] Zou H, Hastie T, Tibshirani R. Sparse principal component analysis [J]. Journal of computational and graphical statistics, 2006, 15(2): 265-286
- [14] d'Aspremont A, El Ghaoui L, Jordan M I, et al. A direct formulation for sparse PCA using semidefinite programming [J]. SIAM review, 2007, 49(3): 434-448
- [15] Journée M, Nesterov Y, Richtárik P, et al. Generalized power method for sparse principal component analysis [J]. The Journal of Machine Learning Research, 2010, 11: 517-553

- [16] Yuan X T, Zhang T. Truncated power method for sparse eigenvalue problems[J]. The Journal of Machine Learning Research, 2013, 14(1): 899-925
- [17] Saad Y. Numerical methods for large eigenvalue problems[M]. Manchester: Manchester University Press, 1992
- [18] Mackey L W. Deflation methods for sparse pca[C]// Advances in Neural Information Processing Systems. 2009: 1017-1024
- [19] Cadima J, Jolliffe I T. Loading and correlations in the interpreta-

tion of principle components[J]. Journal of Applied Statistics, 1995, 22(2): 203-214

- [20] Vines S K. Simple principal components[J]. Journal of the Royal Statistical Society: Series C (Applied Statistics), 2000, 49(4): 441-451
- [21] Jolliffe I T, Trendafilov N T, Uddin M. A modified principal component technique based on the LASSO[J]. Journal of Computational and Graphical Statistics, 2003, 12(3): 531-547

(上接第 437 页)

图 4 所示的环境进行测试和验证。

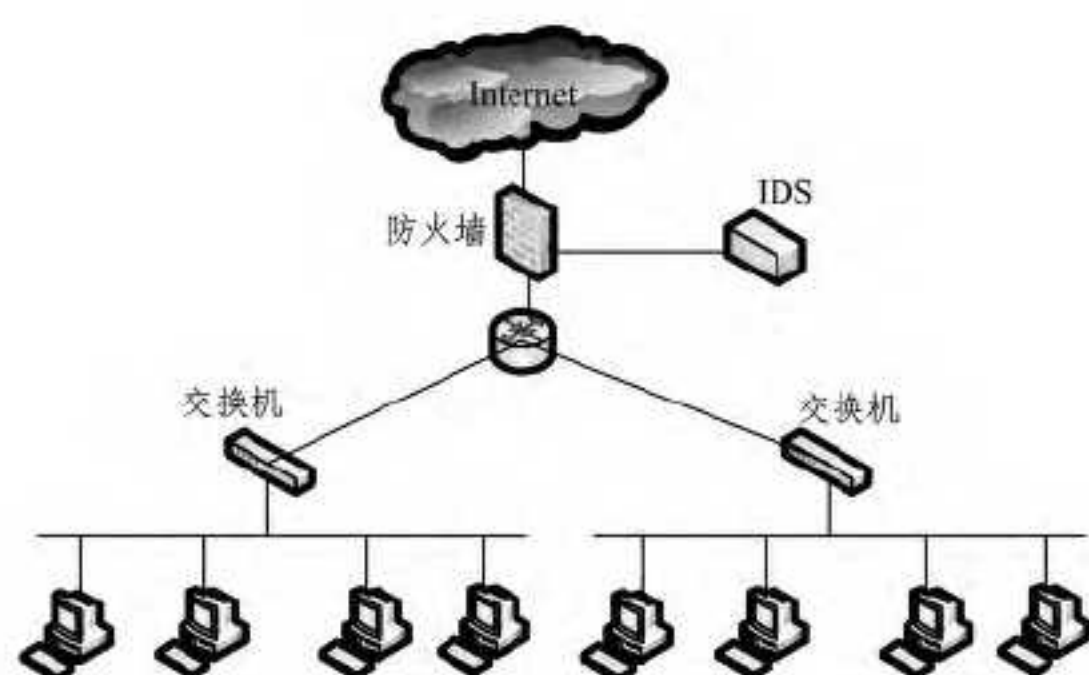


图 4 实验环境图

实验中放置了 10 台配置为 window XP/i3/3G PC 主机连接在互联网中, 防火墙采用天融信千兆防火墙, IDS 为启明

星辰千兆 IDS。

3.2 实验结果

首先对所有样本集中数据生成态势决策表, 提取出大量高质量高可信的规则, 并将规则转换成机器能识别的语言。其次, 对文中涉及的指标进行量化。实验测试环境中开通了 FTP、Telnet、Http、DNS、SNMP 服务, 各服务所占的比 $S_p = \{0.1, 0.25, 0.35, 0.15, 0.15\}$, 按照每台主机上所开服务、拥有资源以及漏洞情况, 将环境中 10 台主机的权重量化为 $H_p = \{0.103, 0.138, 0.172, 0.069, 0.069, 0.138, 0.172, 0.034, 0.172, 0.034, 0.069, 0.034\}$ 。根据实时攻击感知到的攻击结果, 结合式(2)~式(5)以及各变量的量化值计算每时每刻的态势值。态势值越大, 说明网络越处于不安全状态。取 2014 年 9 月 27 日 18:00—20:30 检测到的数据, 计算得到各个时刻的态势值。实验中每 5 分钟评估一次态势值。得到如表 6 所列的态势数据。

表 6 实验结果

时间	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
态势值	845	785	2034	560	1467	365	895	768	803	654	358	295	103	206	489
时间	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
态势值	579	156	267	810	2327	754	796	1450	976	895	1320	1020	846	591	768

根据表 6 的数据绘制得到图 5 所示的态势曲线图。

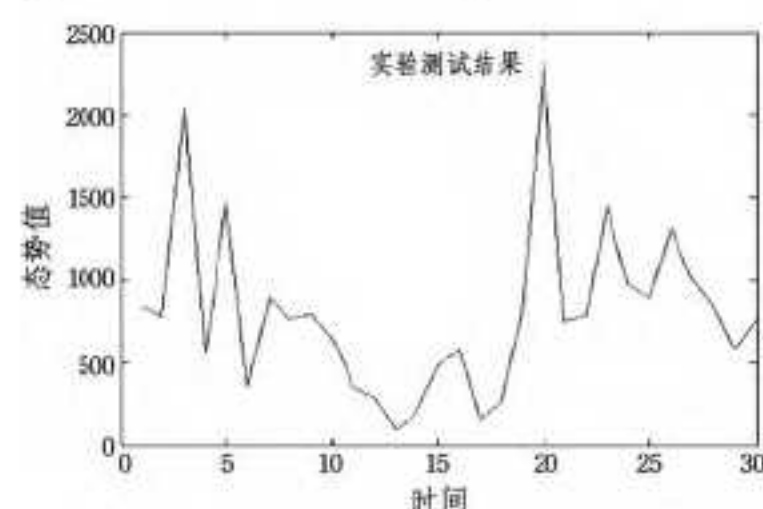


图 5 实验结果

从实验结果可以看出有几次的安全态势值的峰值较高, 网络处于不安全状态, 需要管理员采取一些措施改善网络状况。

结束语 本文提出了基于粗糙集的实时态势评估, 首先用粗糙集中属性约简确定对安全态势产生影响的有效因子; 其次, 采用决策表进行规则的提取和发现。这一过程没有人为因素的干扰, 保证了提取规则集的可靠性、准确性。同时由于需要及时发现复杂攻击, 引入实时攻击检测引擎加载提取到的规则集, 采用实时流计算思想对流经的安全事件数据流在线检测和分析, 并将分析检测的结果作为实时态势评估的依据。实时攻击检测的结果在一定程度上保证了网络安全态势评估的准确性、实时性和客观性。最后经实验验证, 文中态势评估的结果具有较高的实时性和准确性。由于攻击检测引擎不仅要加载规则树到内存, 同时还需要具有一定的存储能力, 因此当规则集数量多而复杂时, 可能需要消耗较多的内

存资源, 因此, 下一步需要对方法进行分布式化, 进行资源的合理分配。

参考文献

- [1] 龚正虎, 卓莹. 网络态势感知研究[J]. 软件学报, 2010, 21(7): 1605-1609
- [2] 陈秀真, 郑庆华, 管晓宏, 等. 层次化网络安全威胁态势量化评估方法[J]. 软件学报, 2006, 17(4): 885-897
- [3] 王娟, 张凤荔, 傅翀, 等. 网络态势感知中的指标体系研究[J]. 计算机应用, 2007, 27(8): 1907-1909
- [4] 卓莹, 何明, 龚正虎. 网络态势评估的粗集分析模型[J]. 计算机工程与科学, 2012, 34(3): 1-5
- [5] 赖积保, 王颖, 王慧强, 等. 基于多源异构传感器的网络安全态势感知系统结构研究[J]. 计算机科学, 2011, 38(3): 144-149
- [6] 石波, 谢小权. 基于 D-S 证据理论的网络安全态势预测方法研究[J]. 计算机工程与设计, 2013, 34(3): 821-825
- [7] 康长青, 郭立红, 罗艳春, 等. 基于模糊贝叶斯网络的态势威胁评估模型[J]. 光电工程, 2008, 35(5): 1-5
- [8] 王琳, 寇英信. Dempster-Shafer 证据理论在空战态势评估方面的应用[J]. 光电与控制, 2007, 14(6): 155-157
- [9] Pawlak Z. Rough Sets[J]. International Journal of Information and Computer Science, 1982, 11(5): 341-356
- [10] Pawlak Z, Gzymala Busse J, Slowinski R. Rough sets[J]. Communications of the ACM, 1995, 38(11): 88-95
- [11] 王国胤, 姚一豫, 于一洪. 粗糙集理论与应用研究综述[J]. 计算机学报, 2009, 32(7): 1229-1246